Naval Research Laboratory
Washington, DC 20375-5000

NRL Memorandum Report 6254

Subroutine Probdif

G. V. TRUNK

Radar Analysis Branch
Radar Division

September 12, 1988

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1a REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b RESTRICTIVE MARKINGS
2a SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b DECLASSIFICATION / DOWNGRADING SCHEDULE	
4 PERFORMING ORGANIZATION REPORT NUMBER(S) NRL Memorandum Report 6254	5 MONITORING ORGANIZATION REPORT NUMBER(S)
6a NAME OF PERFORMING ORGANIZATION Naval Research Laboratory 6b OFFICE SYMBOL (If applicable) Code 5310	7a NAME OF MONITORING ORGANIZATION
6c ADDRESS (City, State, and ZIP Code) Washington, DC 20375-5000	7b ADDRESS (City, State, and ZIP Code)
8a NAME OF FUNDING / SPONSORING ORGANIZATION Naval Sea Systems Command 8b OFFICE SYMBOL (If applicable)	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c ADDRESS (City, State, and ZIP Code) Washington, DC 20362-5101	10 SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO 64573N PROJECT NO S0954 TASK NO WORK UNIT ACCESSION NO

11 TITLE (Include Security Classification)
Subroutine Probdif

12 PERSONAL AUTHOR(S)
Trunk, G.V.

13a TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (Year, Month, Day)	15 PAGE COUNT
Interim	FROM 3/88 TO 5/88	1988 September 12	19

16 SUPPLEMENTARY NOTATION

17 COSATI CODES			18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Distributions Ordered samples

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

This reports describes a technique for determining the probability that one density is a shifted version of another density. We assume nothing about the density and employ a nonparametric procedure on the sample data.

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	21 ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED
22a NAME OF RESPONSIBLE INDIVIDUAL G.V. Trunk	22b TELEPHONE (Include Area Code) (202) 767-2573 22c OFFICE SYMBOL Code 5310

DD Form 1473, JUN 86 Previous editions are obsolete. SECURITY CLASSIFICATION OF THIS PAGE

S/N 0102-LF-014-6603

CONTENTS

SUBROUTINE PROBDIF

INTRODUCTION

Given n independent points $x_1, x_2, \ldots, x_n$ from the density $f_x(\cdot)$ and m independent points $y_1, y_2, \ldots, y_m$ from the density $f_y(\cdot)$, we wish to determine the probability that

$$f_x(Z) = g(Z)$$
$$f_y(Z) = g(Z-\mu) \; .$$

That is, we wish to determine the probability that one density is shifted by μ from the other density. Since we do not want to assume anything about the density $g(\cdot)$, we will employ a nonparametric procedure.

GENERAL DESCRIPTION OF TECHNIQUE

If the x's and y's come from the same density, there are $(m+n)!/n!m!$ possible orderings of the points; and each ordering is equally likely. If there are n' x's smaller than any y and m' y's larger than any x, the probability that the x's and y's come from a different density is equal to the probability of obtaining any sequence that starts with n' or more x's and ends with m' or more y's,

As an example of this procedure, let us assume that there are 3 x samples 1.0, 2.0, and 3.0 and 3 y samples 3.5, 4.5, and 5.5. Given these examples, there are 20 possible orderings for the samples and these orderings are listed in Table 1. The probability that the x points are μ less than the y points is given in Table 2. For $\mu=0$ the sequence has 3 x's followed by 3 y's. Only sequence #1 satisfies this condition and its probability of occurrence is 0.05. Thus, the probability that the samples do not come from the same density is one minus this probability or 0.95. For $\mu=1$, the sequence starts with 2 x's and ends with 2 y's. Sequences #1 and #2 satisfy this condition, their probability of occurrence is 0.10, and the probability that the samples do not come from the same density is 0.90. For $\mu=2$, the sequence starts with an x and ends with a y. Sequences #1, #2, #3, #5, #6, and #8 satisfy this condition, their probability is 0.3, and thus the probability that the samples do not come from the same density is 0.7. When the sequence starts with y's and ends in x's, the probability of occurrence of the sequences equals the probability that the samples do not come from the same density; i.e., one does not have to subtract the sequence probabilities from 1.

When the number of samples is a moderate number, it becomes impractical to enumerate all possible sequences. For instance, if $n=m=15$, the number of

Manuscript approved May 23, 1988.

Table 1 - Possible Sequences

Sequence #	Sequence
1	xxxyyy
2	xxyxyy
3	xxyyxy
4	xxyyyx
5	xyxxyy
6	xyxyxy
7	xyxyyx
8	xyyxxy
9	xyyxyx
10	xyyyxx
11	yxxxyy
12	yxxyxy
13	yxxyyx
14	yxyxxy
15	yxyxyx
16	yxyyxx
17	yyxxxy
18	yyxxyx
19	yyxyxx
20	yyyxxx

Table 2 - Probability that the density of the x points shifted by μ is less than the density of the y points.

μ	Probability
0	.95
1	.90
2	.70
3	.30
4	.10
5	.05

possible sequences is greater than 10^8. Consequently, we will discuss a different technique for calculating the probability of the various sequences without enumerating them. This technique will also handle the cases where the smallest and largest samples come from the same set of points.

SPECIFIC TECHNIQUE

Let us first define NS_x as the number of x's which are smaller than all of the y's and NS_y as the number of y's which are smaller than all of the x's. Similarly, define NL_x as the number of x's which are larger than all of the y's and NL_y as the number of y's which are larger than all of the x's. Note that if $NS_x > 0$, then $NS_y = 0$; and if $NS_y > 0$, then $NS_x = 0$. Similarly, if $NL_x > 0$, then $NL_y = 0$; and if $NL_y > 0$, then $NL_x = 0$. Based on these four values (NS_x, NS_y, NL_x and NL_y), four different cases are possible:

1. $NS_x > 0$ and $NL_y > 0$,
2. $NS_y > 0$ and $NL_x > 0$,
3. $NS_x > 0$ and $NL_x > 0$,
4. $NS_y > 0$ and $NL_y > 0$.

Each of these cases will now be discussed and a formula will be generated to calculate the probability that the x data set comes from a density which has a smaller location parameter than the y density.

Case 1: $NS_x > 0$ and $NL_y > 0$

If $NS_x = n$, that is all the x's are less than all the y's, the probability of obtaining the first x in the sequence is n/(m+n), the probability of obtaining the second x in the sequence is (n-1)/m+n-1), and the probability of the third x is (n-2)/m+n-2), and the probability of the i-th x is (n+1-i)/(m+n+1-i). Consequently, the probability of this sequence is

$$\frac{n}{m+n} \cdot \frac{n-1}{m+n-1} \cdot \frac{n-2}{m+n-2} \cdots \frac{1}{m+1} = \frac{n!m!}{(m+n)!} \tag{1}$$

If $NS_x < n$, let i be in the range $NS_x \leq i < n$; and let i be the exact number of x's which start a sequence. The probability that exactly i x's start a sequence follows from the above argument and equals

$$\frac{n}{m+n} \cdot \frac{n-1}{m+n-1} \cdots \frac{n+1-i}{m+n+1-i} = \frac{n!}{(n-i)!} \quad \frac{(m+n-i)!}{(m+n)!} \quad . \tag{2}$$

Since the sequence starts with exactly i x's, the next sample must be a y; and the probability of this y is

$$\frac{m}{(m+n-i)} \quad . \tag{3}$$

So far we have considered i+1 samples: i samples coming from the x set and 1 sample coming from the y set. Let us now assume that the sequence ends with exactly j y samples. This sequence of y samples must be proceeded by an x, and the probability of this x is

$$\frac{n-i}{m+n-i-1} \quad . \tag{4}$$

Finally the probability of ending with j y samples is

$$\frac{m-1}{m+n-i-2} \cdot \frac{m-2}{m+n-i-3} \cdots \frac{m-j}{m+n-i-j-1} = \frac{(m-1)!}{(m-j-1)!} \frac{(m+n-i-j-2)!}{(m+n-i-2)!} \quad . \tag{5}$$

Thus, the probability P_{ij} of all the sequences that start with exactly i x's and end with exactly j y's is the product of expressions (2), (3), (4), and (5), i.e.,

$$P_{ij} = \frac{n!}{(n-i)!} \frac{(m+n-i)!}{(m+n)!} \frac{m}{(m+n-i)} \frac{n-i}{(m+n-i-1)} \frac{(m-1)!}{(m-j-1)!} \frac{(m+n-i-j-2)!}{(m+n-i-2)!} \quad . \tag{6}$$

Simplifying this expression, one obtains

$$P_{ij} = \frac{n!m!}{(m+n)!} \frac{1}{(n-i-1)!} \frac{(m+n-i-j-2)!}{(m-j-1)!} \quad . \tag{7}$$

For example, if m=n=3, i=2, and j=1, $P_{ij} = 1/20$, which corresponds to sequence #3 in Table 1. If m=n=3, i=1, and j=1; $P_{ij} = 2/20$, which corresponds to sequences #6 and #8 in Table 1. Equation (7) is valid for $NS_x \leq i < n$ and $NL_y \leq j < m$. Thus, the probability of obtaining sequences which start with at least NS_x x's and end with at least NL_y y's can be obtained by summing Eq. (7) over the proper ranges of i and j and adding expression (1) to this sum. Consequently, by subtracting this probability from 1, one obtains the probability that the x points come from a density which has a smaller location parameter that the y density has. The desired probability is

$$P = 1 - \frac{n!m!}{(m+n)!} \left\{ 1 + \sum_{i=NS_x}^{n-1} \frac{1}{(n-i-1)!} \left[\sum_{j=NL_y}^{m-1} \frac{(m+n-i-j-2)!}{(m-j-1)!} \right] \right\} \quad . \tag{8}$$

<u>Case 2: $NS_y > 0$ and $NL_x > 0$</u>

This case is the dual of case 1. If we let $NS_y \leq i < m$ and $NL_x \leq j < n$, the revised Eq. (8) would yield the probability that the y's are less than the x's. The probability that the x's are less than the y's is 1 minus this probability or

$$P = \frac{n!m!}{(n+m)!} \left\{ 1 + \sum_{i=NS_y}^{m-1} \frac{1}{(m-i-1)!} \left[\sum_{j=NL_x}^{n-1} \frac{(m+n-i-j-2)!}{(n-j-1)!} \right] \right\} \quad . \tag{9}$$

<u>Case 3: $NS_x > 0$ and $NL_x > 0$</u>

This case is likely to occur if n >> m. If $NS_x = NL_x$, the desired probability is

$$P = 0.5 \quad . \tag{10}$$

If $NS_x \neq NL_x$, let NS = Minimum (NS_x, NL_x) and let NL = Maximum (NS_x, NL_x). Then, the sequences of interest are those sequences which start with NL or more x's and end with either NS or less x's or any number of y's. The probability that a sequence starts with n x's and ends with m y's is given by expression (1). Let i be in the range $NL \leq i < n$ and let i be the exact number of x's which start a sequence. The probability that exactly i x's start a sequence is given by (2). Since the sequence starts with exactly i x's, the next sample must be a y; and the probability of this y is given by (3). So far we have considered i+1 samples: i samples coming from the x set and 1 sample coming from the y set. Let us now assume that the sequence ends with exactly j x samples. This sequence of x samples must be proceeded by a y, and the probability of this y sample is

$$\frac{m-1}{m+n-i-1} \quad . \tag{11}$$

The probability of ending with j x samples is

$$\frac{n-i}{m+n-i-2} \cdot \frac{n-i-1}{m+n-i-3} \cdots \frac{n-i-j+1}{m+n-i-j-1} = \frac{(n-i)!}{(n-i-j)!} \; \frac{(m+n-i-j-2)!}{(m+n-i-2)!} \quad . \tag{12}$$

The possible values of j are from 1 to N_1 wjere N_1 is the minimum of NS and (n-i). The probability of starting a sequence with exactly i x's and ending it with exactly j x's is the product of (2), (3), (11), and (12)

$$\frac{n!}{(n-i)!} \; \frac{(m+n-i)!}{(m+n)!} \; \frac{m}{(m+n-i)} \; \frac{m-1}{(m+n-i-1)} \; \frac{(n-i)!}{(n-i-j)!} \cdot \frac{(m+n-i-j-2)!}{(m+n-i-2)!} \tag{13}$$

which simplifies to

$$\frac{n!m(m-1)}{(m+n)!} \; \frac{(m+n-i-j-2)!}{(n-i-j)!} \quad . \tag{14}$$

The sequences of interest which start with exactly i x's can also end with j y samples and this probability is given by (7). Combining (7) and (14) and summing over the proper values of i and j, one obtains

$$P' = \frac{n!m!}{(m+n)!} + \sum_{i=NL}^{n-1} \left\{ \frac{n!m(m-1)}{(m+n)!} \sum_{j=1}^{N_1} \frac{(m+n-i-j-2)!}{(n-i-j)!} + \frac{n!m!}{(m+n)!} \; \frac{1}{(n-i-1)!} \sum_{j=1}^{m-1} \frac{(m+n-i-j-2)!}{(m-j-1)!} \right\} \quad . \tag{15}$$

Rearranging terms, one obtains

$$P' = \frac{m!\,n!}{(m+n)!}\left\{1 + \sum_{i=NL}^{n-1}\left[\frac{1}{(m-2)!}\sum_{j=1}^{N_1}\frac{(m+n-i-j-2)!}{(n-i-j)!} + \frac{1}{(n-i-1)!}\sum_{j=1}^{m-1}\frac{(m+n-i-j-2)!}{(m-j-1)!}\right]\right\} . \qquad (16)$$

Finally, if $NS_x > NL_x$

$$P = 1 - P' \quad ; \qquad (17)$$

and if $NL_x > NS_x$

$$P = P' \quad . \qquad (18)$$

Case 4: $NS_y > 0$ and $NL_y > 0$

This case is the dual of case 3. Letting NS = Minimum (NS_y, NL_y), NL = Maximum (NS_y, NL_y), and N_2 be the minimum of NS and (m-i), the probability P' is given by

$$P' = \frac{m!\,n!}{(m+n)!}\left\{1 + \sum_{i=NL}^{m-1}\left[\frac{1}{(n-2)!}\sum_{j=1}^{N_2}\frac{(m+n-i-j-2)!}{(m-i-j)!} + \frac{1}{(m-i-1)!}\sum_{j=1}^{n-1}\frac{(m+n-i-j-2)!}{(n-j-1)!}\right]\right\} . \qquad (19)$$

If $NS_y > NL_y$,

$$P = P' \quad ; \qquad (20)$$

and if $NL_y > NS_y$

$$P = 1 - P' . \qquad (21)$$

SUBROUTINE PROBDIF

Depending on the values of NS_x, NS_y, NL_x, and NL_y, the desired probability is given in Eq. (8), Eq. (9), Eqs. (16-18), or Eqs. (19-21). All of these equations involve double summations over factorials and the computation time will vary as n^3 (or m^3) if these formula are implemented directly. However, since the terms involve factorials, iterative formulas can be developed and the computation time can be made proportional to n^2. Such a technique was incorporated into subroutine PROBDIF which calculates the desired probability.

The calling sequence for the subroutine is

```
CALL PROBDIF (nx,x,ny,y,xdb,prob,iset)
```

where

nx is number of points from the x distribution
x(.) are points from the x distribution
ny is number of points from the y distribution
y(.) are points from the y distribution
xdb is the db difference to be tested
prob is the probability of a xdb difference
iset=1 indicates that points have already been ordered from smallest to largest.

Note, if successive calls are made to PROBDIF with different values of xdb, iset should be set to 1 after the first call to minimize computation time. A Fortran listing of PROBDIF is given in appendix A.

SUMMARY

Given two sets of data, this report describes a technique for determining the probability that the first density is a shifted version of the second density. We assumed nothing about the density and used a nonparametric procedure based on the smallest and largest samples of the combined data set. A computer program was written to calculate the desired probability.

APPENDIX A

```
      subroutine probdif(nx,x,ny,y,xdb,prob,iset)
c
c      calculates probability that x distribution is x dB less
c      than y distribution
c
c      nx is number of points from x distribution
c      x(.) are points from x distribution
c      ny is number of points from y distribution
c      y(.) are points from y distribution
c      xdb is the db difference to be tested
c      prob is the probability of a xdb difference
c      iset=1 indicates that points have already been ordered
c
      dimension x(100),y(100),xo(100),yo(100)
      if (iset.eq.1) go to 70
      n=nx+ny
c
c      ordering of x points
c
      do 5 i=1,nx
  5      xo(i)=x(i)

      if (nx.eq.1) go to 30

      do 20 i=1,nx-1
      do 10 j=i+1,nx
      if (xo(i).lt.xo(j)) go to 10
      t=xo(j)
      xo(j)=xo(i)
      xo(i)=t
 10   continue
 20   continue
c
c      ordering of the y points
c
 30   do 35 i=1,ny
 35      yo(i)=y(i)

      if (ny.eq.1) go to 70
```

```
      do 60 i=1,ny-1
      do 40 j=i+1,ny
      if (yo(i).lt.yo(j)) go to 40
      t=yo(j)
      yo(j)=yo(i)
      yo(i)=t
 40   continue
 60   continue

 70   continue
c
c       nsmallx is the # of x points less than yo(1)-xdb
c       nsmally is the # of y points less than xo(1)-xdb
c
      nsmallx=0
      nsmally=0
      p1=1.0

      do 80 i=1,nx
      if (xo(i).gt.yo(1)-xdb) go to 85
      nsmallx=nsmallx+1
 80   continue

 85   if (nsmallx.eq.0) go to 95
      go to 120

 95   do 100 j=1,ny
      if (xo(1).lt.yo(j)-xdb) go to 105
      nsmally=nsmally+1
 100  continue

 105  continue
c
c       nlargey is the # of y points greater than xo(nx)-xdb
c       nlargex is the # of x points greater than yo(ny(-xdb
c
 120  nlargey=0
      nlargex=0

      do 130 j=1,ny
      jj=ny+1-j
      if (yo(jj).lt.xo(nx)-xdb) go to 135
      nlargey=nlargey+1
 130  continue

 135  if (nlargey.eq.0) go to 145
      go to 170
```

```
  145   do 150 i=1,nx
        ii=nx+1-i
        if (yo(ny).gt.xo(ii)+xdb) go to 155
        nlargex=nlargex+1
  150   continue

  155   continue

  170   continue
c
c        calculation of probability
c
        prob=0.0
        nxp=nx+1
        nyp=ny+1
        np=n+1
        if (nsmallx.gt.0.and.nlargey.gt.0) go to 200
        if (nsmally.gt.0.and.nlargex.gt.0) go to 300
        if (nsmallx.gt.0.and.nlargex.gt.0) go to 400
        if (nsmally.gt.o.and.nlargey.gt.0) go to 500
        print 50
   50   format ('     ****** impossible condition ******')
        return
c
c        case #1  nsmallx>0 and nlargey>0
c
  200   f=1.

        do 250 i=nsmallx,nx
        if (i.gt.nsmallx) go to 210

        do 205 j=1,nsmallx
        nxp=nxp-1
        np=np-1
        f=(f*nxp)/np
  205   continue

        if (i.lt.nx) go to 210
  208   h=1.
        go to 245
  210   if (i.eq.nx) go to 208
        nxp=nxp-1
        np=np-1
        f=(f*nxp)/np
        m=np-1
        nyp=ny
        g=(1.*nyp)/m

        do 240 j=nlargey,ny-1
        if (j.gt.nlargey) go to 230
```

```
      do 225 k=1,nlargey
      nyp=nyp-1
      m=m-1
      g=(g*nyp)/m
      h=g
225   continue

      go to 240
230   nyp=nyp-1
      m=m-1
      g=(g*nyp)/m
      h=h+g
240   continue

245   prob=prob+f*h
250   continue

      prob=1.-prob
      return
c
c      case #2  nsmally>0 and nlargex>0
c
300   continue
      f=1.

      do 350 i=nsmally,ny
      if (i.gt.nsmally) go to 310

      do 305 j=1,nsmally
      nyp=nyp-1
      np=np-1
      f=(f*nyp)/np
305   continue

      if (i.lt.ny) go to 310
308   h=1.
      go to 345
310   if (i.eq.ny) go to 308
      nyp=nyp-1
      np=np-1
      f=(f*nyp)/np
320   m=np-1
      nxp=nx
      g=(1.*nxp)/m

      do 340 j=nlargex,nx-1
      if (j.gt.nlargex) go to 330
```

```
      do 325 k=1,nlargex
      nxp=nxp-1
      m=m-1
      g=(g*nxp)/m
      h=g
325   continue

      go to 340
330   nxp=nxp-1
      m=m-1
      g=(g*nxp)/m
      h=h+g
340   continue

345   prob=prob+f*h
350   continue

      return
c
c      case #3 nsmallx>0 and nlargey>0
c
400   continue
      if (nsmallx.ne.nlargex) go to 401
      prob=0.5
      return
401   ns=min(nsmallx,nlargex)
      nl=max(nsmallx,nlargex)
      f=1.0

      do 450 i=nl,nx
      if (i.gt.nl) go to 410

      do 405 j=1,nl
      nxp=nxp-1
      np=np-1
      f=(f*nxp)/np
405   continue

      if (i.lt.nx) go to 420
408   h=1.
      go to 445
410   nxp=nxp-1
      np=np-1
      f=(f*nxp)/np
      if (i.eq.nx) go to 408
420   h=0.
      m=np-1
      nyp=ny
      g=(1.*nyp)/m
      if (nyp.eq.1) go to 422
      m=m-1
      nyp=nyp-1
      g=(g*nyp)/m
```

```
422   f1=1.0
      nxpp=nxp

      do 425 j=1,ns
      nxpp=nxpp-1
      if (nxpp.eq.0) go to 426
      m=m-1
      f1=(f1*nxpp)/m
      h=h+g*f1
425   continue

426   m=np-1
      g=(nxp-1.)/m
      nyp=ny
      m=m-1
      g=(g*nyp)/m
      f1=1.

      do 440 j=1,ny-1
      nyp=nyp-1
      m=m-1
      f1=(f1*nyp)/m
      h=h+g*f1
440   continue

445   prob=prob+f*h
450   continue

      if (nsmallx.gt.nlargex) prob=1.0-prob
      return
c
c      case #4 nsmally>0 and nlargey>0
c
500   continue
      if (nsmally.ne.nlargey) go to 501
      prob=0.5
      return
501   ns=min(nsmally,nlargey)
      nl=max(nsmally,nlargey)
      f=1.0

      do 550 i=nl,ny
      if (i.gt.nl) go to 510

      do 505 j=1,nl
      nyp=nyp-1
      np=np-1
      f=(f*nyp)/np
505   continue
```

```
      if (i.lt.ny) go to 520
508   h=1.
      go to 545
510   nyp=nyp-1
      np=np-1
      f=(f*nyp)/np
      if (i.eq.ny) go to 508
520   h=0.
      m=np-1
      nxp=nx
      g=(1.*nxp)/m
      if (nxp.eq.1) go to 522
      m=m-1
      nxp=nxp-1
      g=(g*nxp)/m
522   f1=1.0
      nypp=nyp

      do 525 j=1,ns
      nypp=nypp-1
      if (nypp.eq.0) go to 526
      m=m-1
      f1=(f1*nypp)/m
      h=h+g*f1
525   continue

526   m=np-1
      g=(nyp-1.)/m
      nxp=nx
      m=m-1
      g=(g*nxp)/m
      f1=1.

      do 540 j=1,nx-1
      nxp=nxp-1
      m=m-1
      f1=(f1*nxp)/m
      h=h+g*f1
540   continue

545   prob=prob+f*h
550   continue
      if (nsmally.lt.nlargey) prob=1.0-prob
      return
      end
```